



UNIVERSITY OF
ILLINOIS LIBRARY
AT URBANA-CHAMPAIGN
BOOKSTACKS

H THE HECKMAN BINDERY, INC.
or North Manchester, Indiana

V JUST FONT SLOT TITLE

H CC 1W 22 BEER
21 FACULTY
20 WORKING
19 F4PSN

H CC 1W 8 1590
7 NO. 1700-1718

H CC 1W 330
B381 100 >
no. 1705-1718
cop. 2

H CC 7W 47MPEINTS
U. OF ILL.
LIBRARY
URBANA

KEL.

BINDING COPY					
PERIODICAL		CUSTOM		STANDARD	
BOOK		MUSIC		ECONOMY	
				THESIS	
				NO. VOLS. THIS TITLE	
				LEAD ATT.	
ACCOUNT		LIBRARY		NEW	
BOOK		CUSTON		RUB OR SAMPLE	
STX3		COLLATING		TITLE ID	
56672 001 ACCOUNT NAME				AUTH. 1ST	
E01912400 ID #2		NOTES		BINDING FREQUENCY	
STX3 COLLATING				WHEEL	
30				SYS ID	
ADDITIONAL INSTRUCTIONS					
Dept=STX3 Lot#=20 Item#151 HNM=[2X# 1CR2ST3CR MARK BY # B4 91] <i>4/29/</i>					
SEP SHEETS		PTS BD PAPER		TAPE STUBS	
PAPER		BUCK		CLOTH EXT	
POCKETS		SPECIAL PREP			
PAPER		GUM			
CLOTH		FILLER			
LEAF ATTACH.					
INSERT MAT		ACCOUNT LOT NO			
		#20			
PRODUCT TYPE		ACCOUNT PIECE NO			
11		150			
HEIGHT		GROUP CARD		VOL. THIS TITLE	
11		1		160	
COVER SIZE		X 4			
00127752					

STX

330
B385
No. 1716 COPY 2

The Library of the
APR 1 1991

University of Illinois
at Urbana-Champaign

Convergence of Learning Algorithms with Constant Learning Rates

C.-M. Kuan

*Department of Economics
University of Illinois*

K. Hornik

*Tekhnische Universität Wien
Vienna, Austria*



Bureau of Economic and Business Research
College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

Digitized by the Internet Archive
in 2011 with funding from
University of Illinois Urbana-Champaign

<http://www.archive.org/details/convergenceoflea1716kuan>

BEBR

FACULTY WORKING PAPER NO. 90-1716

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

December 1990

Convergence of Learning Algorithms With Constant Learning Rates

C.-M. Kuan
Department of Economics
University of Illinois at Urbana-Champaign

and

K. Hornik
Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien, Vienna, Austria

The first author is grateful for support by the Investors in Business Education and the Bureau of Economic and Business Research of the University of Illinois.

Abstract

We investigate the behavior of neural network learning algorithms with a small, constant learning rate ϵ in stationary, random input environments. It is rigorously established that the sequence of weight estimates can be approximated by a certain ordinary differential equation, in the sense of weak convergence of random processes as ϵ tends to zero. As applications, back-propagation in feedforward architectures and some feature extraction algorithms are studied in more detail.

1 Introduction

For understanding the performance of neural network learning algorithms, it is of fundamental importance to investigate how they behave in stationary random input environments. This analysis yields information about the asymptotic properties of the learned connection weights as the number of training samples increases without bound. Thus far, only algorithms with learning rates tending to zero have been studied. However, most neural network learning is conducted using a small, *constant* learning rate. In this paper, we investigate the limiting behavior of such algorithms.

An on-line (local) learning algorithm can be written as

$$\theta_{n+1} = \theta_n + \eta_n Q(z_n, \theta_n), \quad (1)$$

where θ is the k -dimensional vector of network weights to be learned and its current estimate at time n is denoted by θ_n , z_n is the training pattern presented at time n , η_n is the learning rate employed at time n , and $Q(\cdot, \cdot)$ is a suitable function characteristic of the algorithm.

The key tool in the analysis of the sequence $\{\theta_n\}$ is the so-called interpolated process $(\theta(t), t \geq 0)$, usually defined by

$$\theta(t) = \theta_n, \quad t_n \leq t < t_{n+1}, \quad (2)$$

where

$$t_0 = 0, \quad t_n = \eta_1 + \cdots + \eta_n.$$

(Rather than working with the above piecewise constant interpolation of the sequence $\{\theta_n\}$, one could also use piecewise linear interpolations. The asymptotic properties of these two processes are very similar.)

If η_n tends to zero at a suitable rate, it can be shown that the interpolated process of $\{\theta_n\}$ eventually follows a solution trajectory of a corresponding ODE (ordinary differential equation) with probability one (Ljung, 1977; Kushner & Clark, 1978). This has allowed researchers to draw very valuable conclusions about the convergence behavior of $\{\theta_n\}$, see e.g. Oja (1982), Oja & Karhunen (1985), White (1989), Sanger (1989), Hornik & Kuan (1990), Kuan & White (1990).

However, if $\eta_n \equiv \epsilon$, a small constant, the estimates $\{\theta_n\}$ never stabilize, and the analysis of the interpolated processes cannot be carried out for fixed ϵ . Hence, we are interested in asymptotics where ϵ becomes smaller and smaller.

In section 2 of this paper, we shall utilize a result by Kushner (1984) to establish that as ϵ tends to zero, the corresponding interpolated processes converge weakly to the solution of the associated ODE. The result is then applied to two of the most important network architectures in section 3. The proof of the main theorem is deferred to the appendix.

2 Convergence to the ODE limit

Consider the algorithm (1) with $\eta_n \equiv \epsilon$. In this case, we shall write $\{\theta_n^\epsilon\}$ for the sequence of estimates to emphasize the dependence on ϵ , i.e.

$$\theta_{n+1}^\epsilon = \theta_n^\epsilon + \epsilon Q(z_n, \theta_n^\epsilon). \quad (3)$$

The corresponding interpolated process $(\theta^\epsilon(t), t \geq 0)$ is given by

$$\theta^\epsilon(t) = \theta_n^\epsilon, \quad n\epsilon \leq t < (n+1)\epsilon. \quad (4)$$

Our key result is that for $\epsilon \rightarrow 0$, the interpolated processes can be approximated by the solution of an ODE, in the sense of weak convergence of random processes. We have the following definition.

Definition. Let $\{\xi^\epsilon, \epsilon \geq 0\}$ be a family of random elements with values in some metric space (X, ρ) . We say that ξ^ϵ converges weakly to ξ^0 , symbolically $\xi^\epsilon \Rightarrow \xi^0$, iff

$$\lim_{\epsilon \rightarrow 0} \mathbf{E} f(\xi^\epsilon) = \mathbf{E} f(\xi^0)$$

for all bounded, continuous real functions f on X .

Weak convergence is an extension of the familiar concept of convergence in distribution of sequences of \mathbb{R}^l -valued random variables to families of abstract valued random elements; as a basic reference, we recommend Billingsley (1968). If h is a continuous function from X to \mathbb{R}^l and $\xi^\epsilon \Rightarrow \xi^0$, then $h(\xi^\epsilon) \rightarrow_d h(\xi^0)$, where “ \rightarrow_d ” denotes convergence in distribution. (Billingsley, 1968, page 29).

In our case, we shall regard the interpolated processes $\theta^\epsilon(\cdot)$ as random processes with values in $X = D^k[0, T_\infty)$, the space of all functions from $[0, T_\infty)$ to \mathbb{R}^k which are right continuous with left-hand limits at every $0 \leq t < T_\infty$. Here, T_∞ is the supremum over all T such that the limiting ODE has a unique solution on $[0, T]$ with probability one; in particular, if it has a unique global solution on $[0, \infty)$ for every initial condition, then $T_\infty = \infty$. As a metric on X we shall use

$$\rho(\xi, \eta) = \sum_{m=1}^{\infty} 2^{-m} \min(1, \sup_{0 \leq t \leq T_m} |\xi(t) - \eta(t)|), \quad \xi, \eta \in X,$$

where $\{T_m\}$ is an increasing sequence with $\lim_{m \rightarrow \infty} T_m = T_\infty$, such that X is given the topology of uniform convergence on bounded subintervals of $[0, T_\infty)$.

In order to obtain the ODE limit of $\theta^\epsilon(\cdot)$, we must be able to average out the randomness resulting from the pattern sequence $\{z_n\}$, so that the limiting process (as ϵ tends to zero) eventually follows the “mean” behavior. This is the basic idea of the so-called direct averaging method of Kushner (1984, chapter 5), cf. also Kushner & Shlewart (1984).

We assume the following conditions.

- [A 1] $\{z_n\}$ is a (strictly) stationary and ergodic sequence of random vectors.
- [A 2] For each N , there exists a function $L_N(z)$ such that $L_N(z_n)$ is integrable and
$$\sup_{|\theta|, |\tilde{\theta}| \leq N} |Q(z, \theta) - Q(z, \tilde{\theta})| \leq L_N(z) |\theta - \tilde{\theta}|. \quad (5)$$
- [A 3] For each θ , $Q(z_n, \theta)$ is square integrable with expectation

$$\bar{Q}(\theta) := \mathbf{E} Q(z_n, \theta).$$

These conditions are not the weakest possible, but they can easily be verified or interpreted. The stationarity and ergodicity assumption [A 1] applies when we have time series data; in particular, it is satisfied when the training patterns are identically distributed, independent random variables. [A 2] is a Lipschitz-type of smoothness condition on the function Q , which is satisfied in many interesting neural network applications, as shown in later examples. [A 3] defines the corresponding ODE. Of course, conditions [A 2] and [A 3] are met if the patterns are bounded and Q is continuously differentiable in both arguments.

The following theorem is based on theorem 5.1 in Kushner (1984); its proof is given in the appendix.

Theorem. Assume [A 1] to [A 3] and let $\theta_0^\epsilon \equiv \theta_0$, a fixed vector or a random vector independent of ϵ . Then

$$\theta^\epsilon(\cdot) \Rightarrow \theta(\cdot), \quad (6)$$

where $\theta(\cdot)$ is the solution of the ODE

$$\dot{\theta} = \bar{Q}(\theta), \quad \theta(0) = \theta_0.$$

In particular, if $0 \leq t_1 \leq \dots \leq t_l < T_\infty$,

$$(\theta_{t_1/\epsilon}^\epsilon, \dots, \theta_{t_l/\epsilon}^\epsilon) \rightarrow_{\mathcal{D}} (\theta(t_1), \dots, \theta(t_l)). \quad (7)$$

If in addition θ_0 is nonrandom, then for each $T < T_\infty$,

$$\sup_{0 \leq t \leq T} |\theta^\epsilon(t) - \theta(t)| \rightarrow 0 \quad \text{in probability.} \quad (8)$$

Assuming θ_0 to be nonrandom does not really impose a restriction; in fact, in virtually all neural network applications, the initial θ_0 is, although probably chosen at “random” with the aid of some random number generator, independent from the learning patterns, and we can analyze the behavior of $\{\theta_n^\epsilon\}$ conditional on the initial weights.

The above theorem establishes a very close relation between the sequence of estimates generated by a neural network learning algorithm with small constant learning rates and the solution paths of its associated ODE. Let Θ_d be the set of all “desired” limit points of the algorithm; usually, Θ_d consists of all θ which minimize some criterion function, e.g. the mean square error when approximating targets by actual network outputs in supervised learning. Clearly, we expect the algorithm to perform well if the domain of attraction of Θ_d contains all (or most) reasonable initial configurations. In particular, Θ_d should contain at least one asymptotically stable equilibrium of the ODE. Of course, optimal performance is achieved if there is a single, *globally attractive* equilibrium; however, we are unaware of any neural network learning algorithm which has this property. Taking Error Back-Propagation (EBP) as the most prominent example, it is well known that the error functions for most architectures contain a multiplicity of local minima which are equilibria of the associated ODE.

A global asymptotic analysis of the solution paths of the ODE and in particular, an explicit characterization of the domains of attraction of the equilibria, usually cannot be carried out. Nevertheless, reasonable performance can be expected in the cases where the set of all asymptotically stable equilibria is contained in Θ_d ; some of the feature extraction algorithms investigated in the next section have this property. However, it has to be pointed out that it may already be impossible to give a closed analytic description of the set of all equilibrium points (cf. EBP).

3 Applications

3.1 Error Back-Propagation

Consider the following learning problem. We are given a feedforward network architecture with output $o = F(x, \theta)$; here, x is the network input, θ is the vector of all adjustable weights and F is a function characteristic of the network topology. Suppose that we are also given a sequence of identically distributed training patterns $z_n = (x_n, y_n)$ and that it is desired to adjust the network weights in a way that the mean square approximation error

$$\Phi(\theta) = \frac{1}{2} \mathbf{E} |y_n - F(x_n, \theta)|^2$$

is minimized. The most prominent (on-line) algorithm which has been proposed in the neural network literature is the error back-propagation (EBP) algorithm (Rumelhart, Hinton & Williams, 1986), which is a simple gradient descent algorithm allowing for efficient updating of the weights due to the feedforward architecture. In this case,

$$Q(x, y, \theta) = \nabla F(x, \theta)(y - F(x, \theta)), \quad \bar{Q}(\theta) = \mathbf{E} Q(x_n, y_n, \theta) = -\nabla \Phi(\theta).$$

(The “ ∇ ” symbol denotes taking all partial derivatives with respect to the components of θ .) Hence, if the conditions of our theorem are satisfied, we have $\theta^\epsilon(\cdot) \Rightarrow \theta(\cdot)$, where $\theta(\cdot)$ solves the ODE

$$\dot{\theta} = -\nabla\Phi(\theta).$$

A simple application of the chain rule yields that

$$\frac{d}{dt} \Phi(\theta(t)) = -|\nabla\Phi(\theta(t))|^2,$$

thus Φ is strictly reduced along the solution paths of the ODE unless an equilibrium is reached. If we knew in addition that the level sets $\Theta_N = \{\theta : \Phi(\theta) \leq N\}$ are bounded subsets of \mathbb{R}^k , we could conclude that the set of equilibria is *globally* attractive; however, this condition is not satisfied in many important applications, e.g. in multilayer feedforward architectures with bounded hidden layer activation functions, or in linear architectures with a bottleneck layer as described in Baldi & Hornik (1989). Therefore, although the local minima of Φ are of course the asymptotically stable equilibria of the ODE, it is not necessarily true that the solution paths of the ODE converge to a local minimum of Φ for “most” (e.g. in the sense that the exceptional set has Lebesgue measure zero) initial values, as is very often claimed.

If we assume that the training patterns z_n have bounded fourth moments, conditions [A 2] and [A 3] can easily be verified for the usual multilayer multioutput feedforward architectures with logistic or arctangent hidden layer activation functions. As a (notationally convenient) illustration, consider the following single hidden layer network where the i -th output component o_i is given by

$$o_i = f_i(x, \theta) = g_i \left(\sum_{h=1}^q \beta_{ih} \psi_h \left(\sum_{j=1}^d \alpha_{hj} \xi_j \right) \right), \quad i = 1, \dots, p;$$

here, d , q and p are the numbers of input, hidden and output units, respectively, ξ_j is the j -th input component, and

$$\theta = (\alpha_{11}, \dots, \alpha_{qd}, \beta_{11}, \dots, \beta_{pq}).$$

We then have the following result.

Corollary 1. *Suppose that all activation functions ψ_h and g_i are twice continuously differentiable and that in addition, all hidden layer activation functions ψ_h are bounded and have bounded derivatives up to order two. Then, if the sequence of training patterns $\{z_n\}$ is stationary and ergodic with finite fourth moments,*

$$\theta^\epsilon(\cdot) \Rightarrow \theta(\cdot),$$

where $\theta(\cdot)$ solves the ODE

$$\dot{\theta} = -\nabla\Phi(\theta), \quad \theta(0) = \theta_0.$$

For the proof, let us start by observing that under the above conditions, the network output $F(\mathbf{x}, \theta)$ is uniformly bounded in \mathbf{x} over the weight sets where $|\theta| \leq N$. The nonzero entries in $\nabla F(\mathbf{x}, \theta)$ are of the form

$$\frac{\partial f_i}{\partial \beta_{ih}} = g'_i(\sigma_i) \psi_h(\rho_h), \quad \frac{\partial f_i}{\partial \alpha_{hj}} = g'_i(\sigma_i) \beta_{ih} \psi'_h(\rho_h) \xi_j,$$

where $\rho_h = \sum_j \alpha_{hj} \xi_j$ and $\sigma_i = \sum_h \beta_{ih} \psi_h(\rho_h)$. Hence, we can find a finite constant C_N such that

$$|\nabla F(\mathbf{x}, \theta)| \leq C_N |\mathbf{x}|, \quad |\theta| \leq N.$$

Similarly, it can be shown that there is a finite constant D_N such that all second order partials of F with respect to components of θ can be bounded by $D_N |\mathbf{x}|$, uniformly in \mathbf{x} over $\{|\theta| \leq N\}$.

We conclude that if in addition inputs and output have finite fourth moments, then $Q(\mathbf{x}_n, y_n, \theta)$ is square integrable by Schwarz's inequality and [A 3] is satisfied. If we let $L_N(\mathbf{x}, y) := \sup_{|\theta| \leq N} \nabla Q(\mathbf{x}, y, \theta)$, then (5) is satisfied and, again using the above estimates, we see that $L_N(\mathbf{x}_n, y_n)$ is integrable, whence [A 2].

3.2 Feature Extraction Algorithms for Linear Networks

For many applications it is very important to train networks to be able to extract the main features inherent in high-dimensional input data streams, thereby significantly reducing data dimensionality. Generally speaking, we are looking for functions F which compress a d -dimensional input vector \mathbf{x} into a p -dimensional output vector $y = F(\mathbf{x})$ (where $p < d$ and usually $p \ll d$) such that, in a sense to be made more precise, y contains "as much information about \mathbf{x} as possible".

If we use the mean square error of the best linear estimate of \mathbf{x} given y (the "linear reconstruction error") as a criterion, this leads to a statistical technique known as Principal Component Analysis (PCA), see Bourlard & Kamp (1988), Linsker (1988), Sanger (1989), Baldi & Hornik (1991). In this case, the outputs are of the form $y = W\mathbf{x}$, and the set of all optimal $p \times d$ matrices W can be described as follows. Let $\lambda_1 \geq \dots \geq \lambda_d$ be the eigenvalues of the input covariance matrix $\Sigma = \mathbf{E} \mathbf{x}_n \mathbf{x}'_n$ (in what follows, ' denotes transpose), and assume for simplicity that the inputs are centered, i.e. $\mathbf{E} \mathbf{x}_n = 0$, and that all eigenvalues are distinct and positive. For $i = 1, \dots, d$, let u_i be a unit length eigenvector of Σ corresponding to the eigenvalue λ_i . Then W is optimal iff its rows span the same p -dimensional subspace of \mathbb{R}^d as u_1, \dots, u_p , i.e. iff $W = R U'_p$, where R is an invertible $p \times p$ matrix and $U_p = [u_1, \dots, u_p]$.

One class of PCA learning algorithms which have been proposed in the literature can be described as follows, see e.g. Baldi & Hornik (1991), Hornik & Kuan (1990). W is decomposed as $W = M A$, where M is an $p \times p$ matrix with

all diagonal entries equal to one. In particular, we could have $M \equiv I$, the $p \times p$ unit matrix. The algorithm is

$$\begin{aligned} A_{n+1}^\epsilon &= A_n^\epsilon + \epsilon Q_A(x_n, A_n^\epsilon, M_n^\epsilon), \\ M_{n+1}^\epsilon &= M_n^\epsilon + \epsilon Q_M(x_n, A_n^\epsilon, M_n^\epsilon), \end{aligned}$$

with $y = Wx = MAx$ and

$$\begin{aligned} Q_A(x, A, M) &= yx' - \Omega_A(yy')A, \\ Q_M(x, A, M) &= \Omega_M(yy'); \end{aligned}$$

both Ω_A and Ω_M are linear operators on the space of $p \times p$ matrices.

The following result follows immediately from our main theorem.

Corollary 2. *Suppose that the starting values A_0 and M_0 are independent from ϵ and that the input sequence $\{x_n\}$ is stationary and ergodic with finite fourth moments. Then*

$$(A^\epsilon(\cdot), M^\epsilon(\cdot)) \Rightarrow (A(\cdot), M(\cdot)),$$

where $(A(\cdot), M(\cdot))$ is the solution of the ODE

$$\begin{aligned} \dot{A} &= MA\Sigma - \Omega_A(MA\Sigma A'M')A, & A(0) &= A_0, \\ \dot{M} &= \Omega_M(MA\Sigma A'M'), & M(0) &= M_0. \end{aligned}$$

If we take Ω_A as the identity mapping and $M \equiv I$, we obtain an algorithm introduced independently by Williams (1985) as the SEC (symmetric error correction) algorithm, by Baldi (1988) as a symmetric simplification of the BP algorithm for a linear d -p-d architecture in autoassociative mode, and by Oja (1989) as the subspace algorithm; for more details, see Baldi & Hornik (1991). Of course, this algorithm is a generalization of the one-unit algorithm introduced in Oja (1982) as a first order approximation to normalized hebbian learning with small learning rates. In this case, the limiting ODE is

$$\dot{A} = A\Sigma - A\Sigma A'A.$$

For the one-unit case (i.e. $p = 1$), the asymptotic behavior of the solutions of this ODE is completely analyzed in Oja & Karhunen (1985). It can be shown that the solution paths always converge to $\pm u_1$ unless the starting value is perpendicular to u_1 .

For $p > 1$, similar global results do not seem to be available. It can be shown that all full rank equilibrium points of the ODE are of the form $A = R[u_{i_1}, \dots, u_{i_p}]'$, where $1 \leq i_1 < \dots < i_p \leq d$ and R is an orthogonal $p \times p$ matrix (see e.g. Baldi & Hornik, 1991). Therefore, as these equilibrium points are not isolated, they cannot be asymptotically stable. More precisely, Krogh & Hertz (1990) show that all equilibria with $\{i_1, \dots, i_p\} \neq \{1, \dots, p\}$ are unstable

and that for equilibria of the form $A = RU_p'$, only the components of small perturbations about the equilibrium A which are perpendicular to the row space of A die out asymptotically. Thus, one might expect that the estimates more or less “randomly” walk around the manifold

$$\mathcal{A} = \{A = RU_p' : R \text{ orthogonal}\}$$

rather than being attracted by one particular equilibrium.

These stability problems disappear if instead we use the asymmetric algorithm introduced in Sanger (1989) as the GHA (generalized hebbian algorithm). In this case, we take $M \equiv I$ and Ω_A as the “lower” operator which sets all entries of an $p \times p$ matrix which are above the main diagonal to zero. As shown in Sanger (1989), see also Hornik & Kuan (1990), the asymptotically stable equilibria of the associated ODE

$$\dot{A} = A\Sigma - \text{lower}(A\Sigma A') A$$

are given by

$$A = [\pm u_1, \dots, \pm u_p]'$$

Therefore, the performance of the GHA should be as good as the one of Oja’s one-unit algorithm (which is of course the GHA for $p = 1$), and it should be superior to the symmetric algorithm. A satisfactory global analysis of the asymptotic behavior of the solution paths of the above ODE has not been carried out thus far. Sanger (1989, p. 463) claims that the domain of attraction of the set of asymptotically stable equilibria consists of all matrices A , which is not true, due to the existence of equilibria which are not asymptotically stable. In fact, it is easily seen that if the rows of the initial $A(0)$ are perpendicular to some u_i , then the same is true for all $A(t)$, $t \geq 0$.

Rubner & Tavian (1990) introduced an algorithm where, upon presentation of a new pattern x , A is updated according to a hebbian learning rule with columnwise normalization, and M is modified using an asymmetric (i.e. hierarchical) decorrelation filter. If instead we use Oja’s one-unit algorithm for each of the rows of A (Baldi & Hornik, 1991), we obtain another algorithm contained in our general class, with the choices $\Omega_A = \text{diag}$ and $\Omega_M = -\text{subdiag}$ (“diag” respectively “subdiag” are the linear operators which set the offdiagonal respectively the superdiagonal entries of a square matrix to zero). The corresponding ODE is

$$\begin{aligned}\dot{A} &= MA\Sigma - \text{diag}(MA\Sigma A'M') A \\ \dot{M} &= -\text{subdiag}(MA\Sigma A'M')\end{aligned}$$

with the appropriate initial conditions; usually, A_0 is “random” and $M_0 = I$. Hornik & Kuan (1990) show that the asymptotically stable equilibria of this ODE are given by

$$A = [\pm u_1, \dots, \pm u_p]', \quad M = I.$$

The global asymptotic behavior of the solution paths has not been described thus far.

Of course, many other PCA learning algorithms exist. Földiák (1989) suggested an algorithm which combines Oja's one-unit algorithm and lateral inhibition terms. As this algorithm uses a feedback rather than the above feedforward architecture, it cannot be dealt with in our framework, because the learning patterns z_n then consist of the new input x_n and some feedback term y_n which depends on all previous inputs and weight estimates (the case of "state dependent noise"). Hornik & Kuan (1990) analyze the asymptotic behavior of such feedback feature extraction algorithms for the case where the learning rates tend to zero at a suitable rate; the case of constant learning rates is currently being investigated.

Appendix – Proof of the theorem

We proceed along the lines of theorem 5.1 in Kushner (1984). (Our notation is different from Kushner's; the correspondencies are $\theta \mapsto x$, $z \mapsto \xi$, and $Q \mapsto G$; also observe that in our case, both Q and $\{z_j\}$ do not depend upon ϵ .)

Due to stationarity, establishing uniform integrability of

$$\left\{ \sup_{|\theta| \leq N} |Q(z_j, \theta)|, j \geq 0 \right\}$$

reduces to showing that $\mathbf{E} \sup_{|\theta| \leq N} |Q(z_j, \theta)| < \infty$ (cf. Billingsley (1968, p. 32)), which follows from

$$\begin{aligned} \sup_{|\theta| \leq N} |Q(z_j, \theta)| &\leq \sup_{|\theta| \leq N} |Q(z_j, \theta) - Q(z_j, 0)| + |Q(z_j, 0)| \\ &\leq \sup_{|\theta| \leq N} L_N(z_j) |\theta| + |Q(z_j, 0)| \\ &\leq N L_N(z_j) + |Q(z_j, 0)|, \end{aligned}$$

integrability of $L_N(z_j)$ and square integrability of $Q(z_j, 0)$, thereby establishing (A 5.2.1).

For $|\theta|, |\tilde{\theta}| \leq N$ we have

$$\mathbf{E} \sup_{|\tilde{\theta} - \theta| \leq \delta} |Q(z_j, \theta) - Q(z_j, \tilde{\theta})| \leq \delta \mathbf{E} L_N(z_j),$$

hence the left hand side tends to zero as $\delta \rightarrow 0$, establishing (A 5.2.2.a). Finally, stationarity and ergodicity of $\{z_n\}$ ensure that

$$\frac{1}{n-m} \sum_{j=m}^{n-1} Q(z_j, \theta) \rightarrow \mathbf{E} Q(z_j, \theta) = \bar{Q}(\theta)$$

in mean square as $n-m \rightarrow \infty$ (Doob, 1953, theorem X.6.1), which in turn implies (A 5.2.3.a).

Theorem 5.1 in Kushner (1984) now yields that if X is given the Skorohod topology (see Kushner, 1984, pages 30–33), then $\theta^\epsilon(\cdot) \Rightarrow \theta(\cdot)$. In fact, Kushner assumes that $T_\infty = \infty$; however, it is straightforward to see that everything goes through *mutatis mutandis* if $T_\infty < \infty$. Proceeding along the lines of chapter 18 in Billingsley (1968), it can be shown that we also have $\theta^\epsilon(\cdot) \Rightarrow \theta(\cdot)$ if X is given the topology of uniform convergence on bounded subintervals.

The remaining assertions can now easily be established. The mappings $x \in X \mapsto (x(t_1), \dots, x(t_l))$ and $x \in X \mapsto \sup_{0 \leq t \leq T} |x(t) - y(t)|$ for fixed (and nonrandom) $y \in X$ are continuous mappings from X to \mathbb{R}^d respectively \mathbb{R} . Using the Continuous Mapping Theorem (Billingsley, 1968, theorem 5.1), (7) follows immediately and (8) together with the fact that weak convergence to a nonrandom limit is equivalent to convergence in probability to that limit, see e.g. Billingsley (1968, p. 25).

References

Baldi, P. (1988). Linear learning: landscapes and algorithms. In Touretzky, D. S. (ed.), *Advances in Neural Information Processing Systems I*, Proceedings of the 1988 NIPS Conference, Denver. Morgan Kaufmann.

Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: learning from examples without local minima. *Neural Networks*, **2**, 53–58.

Baldi, P., & Hornik, K. (1991). Back-propagation and unsupervised learning in linear networks. In Chauvin, Y., and Rumelhart, D. E. (eds.), *Back Propagation: Theory, Architectures and Applications*. Earlbaum Associates.

Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.

Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**, 291–294.

Doob, J. L. (1953). *Stochastic Processes*. New York: Wiley.

Földiák, P. (1989). Adaptive network for optimal linear feature extraction. In *Proceedings of the International Joint Conference on Neural Networks* (pp. I: 401–405). San Diego: SOS Printing.

Hornik, K., & Kuan, C.-M. (1990). *Convergence analysis of local feature extraction algorithms*. Preprint.

Krogh, A., & Hertz, J. A. (1990). Hebbian learning of principal components. In Eckmiller, R., Hartmann, G., and Hauske, G. (eds.), *Parallel Processing in Neural Systems and Computers* (pp. 183–186). Elsevier Science Publishers B.V. (North-Holland).

Kuan, C.-M., & White, H. (1990). *Recursive M-estimation, nonlinear regression and neural network learning with dependent observations*. BEBR Working Paper 90-1703, College of Commerce, University of Illinois, Urbana-Champaign.

Kushner, H. J. (1984). *Approximation and weak convergence methods for random processes*. Cambridge: MIT Press.

Kushner, H. J., & Clark, D. S. (1978). *Stochastic approximation methods for constrained and unconstrained systems*. New York: Springer Verlag.

Kushner, H. J., & Shwartz, A. (1984). Weak convergence and asymptotic properties of adaptive filters with constant gains. *IEEE Transactions on Information Theory*, **IT-30**, 177–182.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, **21**, 105–117.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, **AC-22**, 551–575.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematics and Biology*, **15**, 267–273.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, **1**, 61–68.

Oja, E., & Karhunen, J. (1985). On stochastic approximation of the eigenvectors and the eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, **106**, 69–84.

Rubner, J. & Tavian, P. (1990). *A self-organizing network for principal component analysis*. Preprint, Physics Department, Technische Universität München, Munich, Germany.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Vol. 1 (chap. 8, pp. 318–362). Cambridge, MA: MIT Press.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2, 459–473.

White, H. (1989). Some asymptotic results for learning in single hidden-layer feed-forward network models. *Journal of the American Statistical Association*, 84, 1003–1013.

Williams, R. J. (1985). *Feature discovery through error-correction learning*. Technical Report 8501, Institute of Cognitive Science, University of California, San Diego.



HECKMAN
BINDERY INC.

JUN 95

ound-To-Pics® N. MANCHESTER,
INDIANA 46962

UNIVERSITY OF ILLINOIS URBANA



3 0112 060295919